STAT-615 Project I: Data Analysis

library(tidyverse)

# Introduction

#Load Data  
  
"britain\_species.dat" %>%  
 read\_table(col\_names = FALSE) %>%  
 rename(island = X1,  
 area = X2,  
 elevation = X3,  
 soil = X4,  
 latitude = X5,  
 distance = X6,  
 species = X7) %>%  
 mutate(area = as.double(area),  
 elevation = as.integer(elevation),  
 soil = as.integer(soil),  
 latitude = as.double(latitude),  
 distance = as.double(distance),  
 species = as.integer(species)) -> bird\_datv2  
  
  
bird\_datv2

## # A tibble: 42 x 7  
## island area elevation soil latitude distance species  
## <chr> <dbl> <int> <int> <dbl> <dbl> <int>  
## 1 Ailsa 0.8 340 1 55.3 14 75  
## 2 Anglesey 712. 127 3 53.3 0.2 855  
## 3 Arran 429. 874 4 55.6 5.2 577  
## 4 Barra 18.4 384 2 57 77.4 409  
## 5 Bressay 31.1 226 1 60.1 202. 177  
## 6 Britain 229850. 1343 16 54.3 0 1666  
## 7 Canna 12.7 210 1 57.1 40.6 300  
## 8 Coll 74.1 103 3 56.6 14.5 443  
## 9 Colonsay 44.8 143 1 56.1 31.1 482  
## 10 Eigg 29 393 1 56.9 12.3 453  
## # ... with 32 more rows

For our data analysis project, we were interested in exploring the diversity of species. The importance of this question subject is nontrivial, as preserving species diversity is incredibly important to prevent extinction of them. By finding out what factors are related or reponsible for increased diversity, we can obtain general awareness of them, as well as understand their role in diversity and how to manage them properly to preserve variety of species.

Because this is a very broad subject, we decided to narrow in on a particular dataset. The University of Florida contains a [data repository](http://users.stat.ufl.edu/~winner/datasets.html) which contained a [dataset](http://users.stat.ufl.edu/~winner/data/britain_species.dat) which included information about bird species diversity in the islands, as well as mainland Britain.

The following variables are found in this dataset: - island: name of the island - area: measured in squared kilometers - elevation: highest peak, measured in meters - soil: number of different soil types - latitude - distance: from mainland britain - species: total number of bird species

head(bird\_datv2, 5)

## # A tibble: 5 x 7  
## island area elevation soil latitude distance species  
## <chr> <dbl> <int> <int> <dbl> <dbl> <int>  
## 1 Ailsa 0.8 340 1 55.3 14 75  
## 2 Anglesey 712. 127 3 53.3 0.2 855  
## 3 Arran 429. 874 4 55.6 5.2 577  
## 4 Barra 18.4 384 2 57 77.4 409  
## 5 Bressay 31.1 226 1 60.1 202. 177

ncol(bird\_datv2)

## [1] 7

nrow(bird\_datv2)

## [1] 42

The dataset contains 7 variables, and 42 observations. Although there is no missing data it should be noted that a limitation of this dataset and following analyses is that it is not include a particularly large number of observations, particularly with respect to the number of variables considered.

#Limitations

Since latitude is not particularly useful, it will not be used in any analyses. This will place the observations to variable ratio at 42:6. However this would only be the case for conisdered models in which all other 6 variables are included. Some limitations of the small dataset are that it does not produce strong statistical power, thus, any conclusions, null or alternative, should be interpreted with caution.

#Preliminary Data Analysis

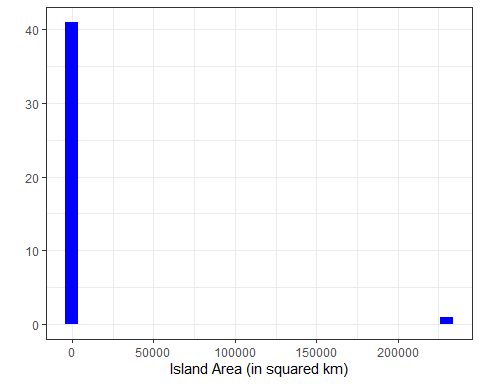
summary(bird\_datv2)

## island area elevation soil   
## Length:42 Min. : 0.50 Min. : 15.0 Min. : 1.000   
## Class :character 1st Qu.: 14.00 1st Qu.: 129.5 1st Qu.: 1.000   
## Mode :character Median : 58.15 Median : 232.0 Median : 2.000   
## Mean : 5740.12 Mean : 355.0 Mean : 2.643   
## 3rd Qu.: 380.38 3rd Qu.: 470.2 3rd Qu.: 3.000   
## Max. :229849.80 Max. :1343.0 Max. :16.000   
## latitude distance species   
## Min. :50.70 Min. : 0.000 Min. : 62.0   
## 1st Qu.:56.12 1st Qu.: 9.175 1st Qu.: 174.8   
## Median :57.05 Median : 30.050 Median : 383.0   
## Mean :57.28 Mean : 63.157 Mean : 392.2   
## 3rd Qu.:59.25 3rd Qu.: 81.075 3rd Qu.: 465.8   
## Max. :60.80 Max. :258.100 Max. :1666.0

#Variable Distributions

##Island Area

ggplot(bird\_datv2, aes(x = area)) +  
 geom\_histogram(fill = "blue") +  
 theme\_bw() +  
 ylab("") +  
 xlab("Island Area (in squared km)") -> area1  
area1



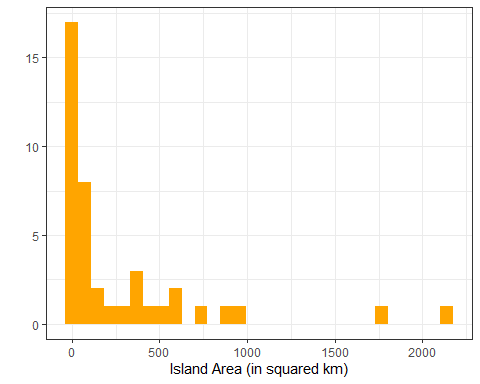
We note that for area there is a clear and extreme outlier.

bird\_datv2 %>%  
 filter(area > 200000)

## # A tibble: 1 x 7  
## island area elevation soil latitude distance species  
## <chr> <dbl> <int> <int> <dbl> <dbl> <int>  
## 1 Britain 229850. 1343 16 54.3 0 1666

Looking into the data we note that this outlier is clearly mainland Britain. To observe a more informative histogram, it is produced without this outlier.

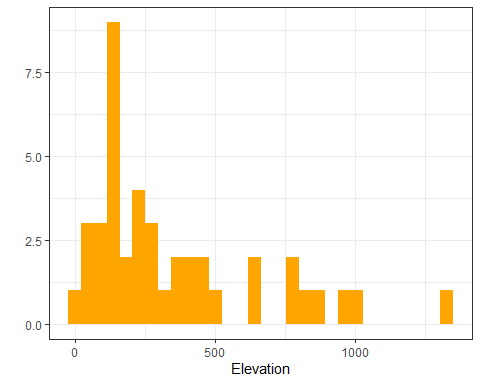
bird\_datv2 %>%  
 filter(area < 200000) %>%  
 ggplot(aes(x = area)) +  
 geom\_histogram(fill = "orange") +  
 theme\_bw() +  
 ylab("") +  
 xlab("Island Area (in squared km)") -> area2  
area2



After exclusion of other the outlier, we note that there are still a few more towards higher areas, and this variable in general appears to be strongly right skewed. Tranformation of this variable might be nessesary. We also note for further data visualization that the observation of mainland Britain will likely be an outlier as well.

##Elevation

ggplot(bird\_datv2, aes(x = elevation)) +  
 geom\_histogram(fill = "orange") +  
 theme\_bw() +  
 ylab("") +  
 xlab("Elevation") -> elev  
elev



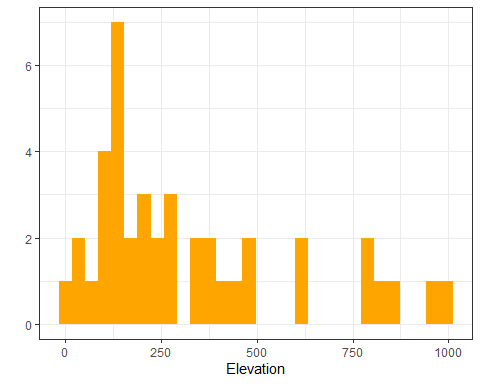
As clearly visible in the histogram above, there is a presence of an extreme outlier which could influence the predictions.

bird\_datv2%>%  
 filter(elevation>1000)

## # A tibble: 2 x 7  
## island area elevation soil latitude distance species  
## <chr> <dbl> <int> <int> <dbl> <dbl> <int>  
## 1 Britain 229850. 1343 16 54.3 0 1666  
## 2 Skye 1735. 1009 5 57.3 0.6 594

The table above denotes the outliers within this variable.

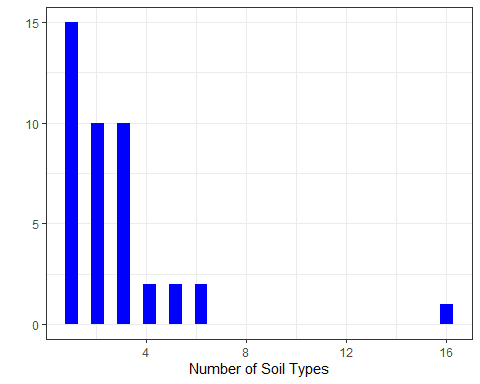
bird\_datv2 %>%  
 filter(elevation < 1010) %>%  
 ggplot(aes(x = elevation)) +  
 geom\_histogram(fill = "orange") +  
 theme\_bw() +  
 ylab("") +  
 xlab("Elevation")



Upon rmeoving the outliers, there is still a few present towards higher elevations. However this variable seems to be skewed towrds the rigth as well. Data manipilation will be required to get accurate estimates from this dataste.

## Number of Soil Types

ggplot(bird\_datv2, aes(x = soil)) +  
 geom\_histogram(fill = "blue") +  
 theme\_bw() +  
 ylab("") +  
 xlab("Number of Soil Types") -> soil  
soil

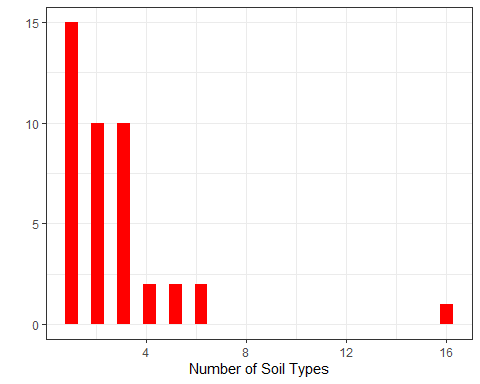
 As clearly visible, there are 3 clear outliers present

bird\_datv2%>%  
 filter(soil>20)

## # A tibble: 0 x 7  
## # ... with 7 variables: island <chr>, area <dbl>, elevation <int>, soil <int>,  
## # latitude <dbl>, distance <dbl>, species <int>

The table above denotes the extreme outliers within this variable.

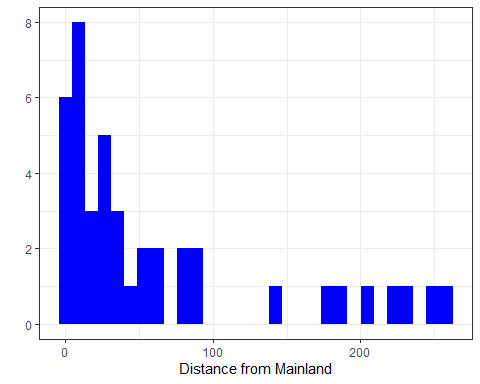
bird\_datv2 %>%  
 filter(soil < 100) %>%  
 ggplot(aes(x = soil)) +  
 geom\_histogram(fill = "red") +  
 theme\_bw() +  
 ylab("") +  
 xlab("Number of Soil Types")



Despite getting rid of the extreme outliers, there is still presence of a few on the right. As other variables before, even this this variable is not normally disctributed and is skewed towards the right which will imoact prediction accuracies.

## Distance from Mainland

ggplot(bird\_datv2, aes(x = distance)) +  
 geom\_histogram(fill = "blue") +  
 theme\_bw() +  
 ylab("") +  
 xlab("Distance from Mainland") -> dis  
dis

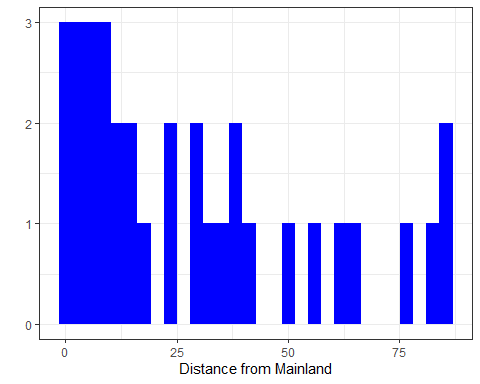
 As clearly visible, there are a few outliers over 100 km .

bird\_datv2%>%  
 filter(distance>100)

## # A tibble: 8 x 7  
## island area elevation soil latitude distance species  
## <chr> <dbl> <int> <int> <dbl> <dbl> <int>  
## 1 Bressay 31.1 226 1 60.1 202. 177  
## 2 Fair 5.2 217 1 59.5 144. 174  
## 3 Fetlar 40.9 159 2 60.6 247. 189  
## 4 Foula 13.5 418 1 60.1 177. 149  
## 5 Shetland 984. 450 6 60.3 189. 421  
## 6 Unst 121. 285 2 60.8 258. 246  
## 7 Whalsay 19.7 120 1 60.4 221 158  
## 8 Yell 217. 205 2 60.6 236. 161

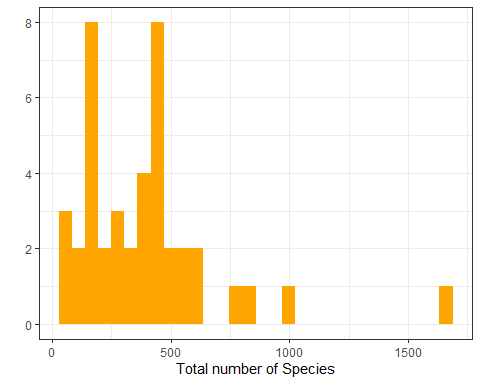
The table above denotes the extreme outliers within this variable.

bird\_datv2 %>%  
 filter(distance < 100) %>%  
 ggplot(aes(x = distance)) +  
 geom\_histogram(fill = "blue") +  
 theme\_bw() +  
 ylab("") +  
 xlab("Distance from Mainland")

 After removing teh outliers, we can see that the distribution is somewhat normal with some skewness towards the right. Slight data transformation will be required.

## Distance from Mainland

ggplot(bird\_datv2, aes(x = species)) +  
 geom\_histogram(fill = "orange") +  
 theme\_bw() +  
 ylab("") +  
 xlab("Total number of Species") -> spe  
spe

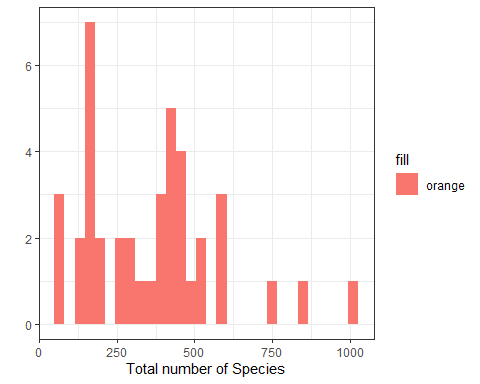
 As clearly visible, there are a few outliers over 100 km .

bird\_datv2%>%  
 filter(species>1500)

## # A tibble: 1 x 7  
## island area elevation soil latitude distance species  
## <chr> <dbl> <int> <int> <dbl> <dbl> <int>  
## 1 Britain 229850. 1343 16 54.3 0 1666

The table above denotes the extreme outliers within this variable.

bird\_datv2 %>%  
 filter(species < 1500) %>%  
 ggplot(aes(x = species,  
 fill = "orange")) +  
 geom\_histogram() +  
 theme\_bw() +  
 ylab("") +  
 xlab("Total number of Species")

 After removing teh outliers, we can see that the distribution is somewhat normal with some skewness towards the right and left. After soem data tranfromation, we will be able to being this normal distribution to be used in a prediction model.

# Corelation Matrix

bird\_datv2\_cor<-bird\_datv2%>%  
 select(area,elevation,soil,distance,species)  
bird\_datv2\_cor

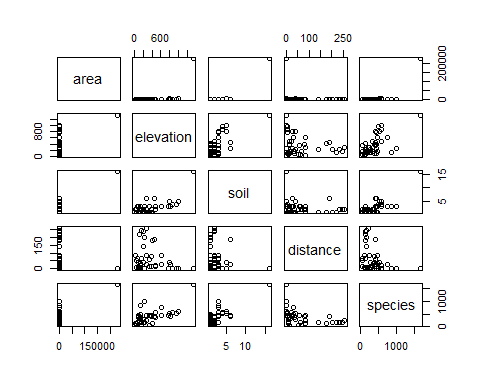
## # A tibble: 42 x 5  
## area elevation soil distance species  
## <dbl> <int> <int> <dbl> <int>  
## 1 0.8 340 1 14 75  
## 2 712. 127 3 0.2 855  
## 3 429. 874 4 5.2 577  
## 4 18.4 384 2 77.4 409  
## 5 31.1 226 1 202. 177  
## 6 229850. 1343 16 0 1666  
## 7 12.7 210 1 40.6 300  
## 8 74.1 103 3 14.5 443  
## 9 44.8 143 1 31.1 482  
## 10 29 393 1 12.3 453  
## # ... with 32 more rows

cor(bird\_datv2\_cor, use = "complete.obs")

## area elevation soil distance species  
## area 1.0000000 0.5079359 0.8411522 -0.1311322 0.6954160  
## elevation 0.5079359 1.0000000 0.6935066 -0.2339001 0.6256577  
## soil 0.8411522 0.6935066 1.0000000 -0.1769162 0.7740322  
## distance -0.1311322 -0.2339001 -0.1769162 1.0000000 -0.4003975  
## species 0.6954160 0.6256577 0.7740322 -0.4003975 1.0000000

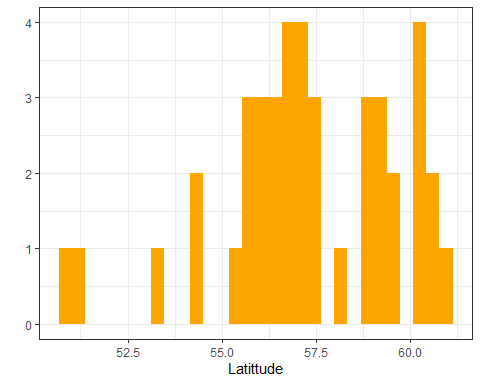
#Scatter Plots

pairs(bird\_datv2\_cor)

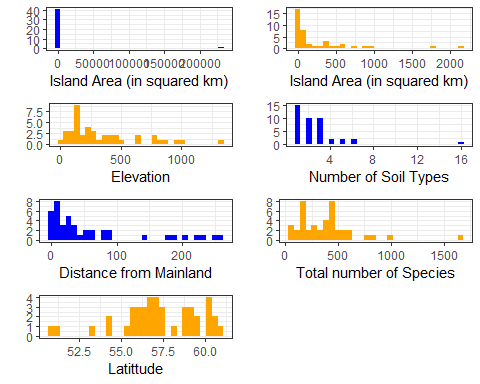


Summary

ggplot(bird\_datv2, aes(x = latitude)) +  
 geom\_histogram(fill = "orange") +  
 theme\_bw() +  
 ylab("") +  
 xlab("Latittude") -> lat  
lat



library(ggpubr)  
ggarrange(area1, area2, elev, soil, dis, spe, lat,  
 ncol = 2, nrow = 4)



library(GGally)  
bird\_datv2 %>%  
 select(species, area, elevation, distance, soil, latitude) %>%  
 ggpairs()

